GENOMIC NETWORK TOMOGRAPHY

Michael G. Rabbat[†]*, Mário A.T. Figueiredo*[‡]*, and Robert D. Nowak*[♯]

[†]McGill University
Electrical and Computer Eng.
Montréal, Québec, CANADA

[‡]Instituto Superior Técnico
Instituto de Telecomunicações
Lisboa, PORTUGAL

[♯]University of Wisconsin-Madison
Electrical and Computer Eng.
Madison, Wisconsin, USA

ABSTRACT

This paper considers the problem of learning cellular signaling networks from incomplete measurements of pathway activity. Cells respond to environmental changes (e.g., starvation, heat shock) via a sequence of intracellular protein-protein interactions, leading to the production of proteins which modify their fundamental operations. Biologists have discovered some of these *signaling pathways*, but the knowledge of cellular signaling is still very incomplete. Mathematically, the problem of *genomic network tomography* (GNT) – identifying cellular signaling networks from biological data – is similar to network inference problems arising in communication systems. This paper formulates GNT and presents a solution which builds on state-of-the-art communication network inference techniques while taking into account uncertainties which are inherent in biological data.

Index Terms— Networks, Biological systems, Communication systems, Monte Carlo methods

1. INTRODUCTION

In living cells, *signaling pathways* communicate information about extracellular conditions from the cell wall to the nucleus, leading to changes in the expression of genes and their protein products, enabling the cell to adapt to its environment. Each signaling pathway comprises a sequence of protein-protein interactions. Viewed collectively, the signaling pathways overlap to form a *signaling network*. Biologists have discovered certain pathways, but the current knowledge of cellular communications is still very incomplete. Due to the intrinsic difficulties of intracellular measurement, we consider the problem of inferring a cellular signaling network from co-occurrence data: observations in the form of lists of the proteins co-occurring in different signaling pathways, without information about pathway structure (*i.e.*, the order of proteins in each pathway). We call this problem *genomic network tomography* (GNT), due to its strong resemblance (both physical and mathematical) to tomographic imaging and network tomography problems.

Formally, the GNT problem is defined as follows. The network is modelled as a directed graph, $G = (V, E)$, with vertex set $V = \{1, 2, \ldots, |V|\}$ and edge set $E \subseteq V \times V$, where $(i, j) \in E$ denotes a directed edge from i to j. The graph contains one vertex for each signaling protein and one edge for each protein-protein interaction. Our goal is to infer the graph structure from a collection of M protein lists, $\mathcal{Y} = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}\}$, where $\mathbf{y}^{(m)} = \{y_1^{(m)}, \ldots, y_{|V|}^{(m)}\} \in \{0, 1\}^{|V|}$, and $y_i^{(m)}$ is a noisy (*i.e.*, possibly wrong) indicator of the observed occurrence ($y_i^{(m)} = 1$) or absence ($y_i^{(m)} = 0$) of protein i in he mth pathway. The aim of GNT is to infer E from $\mathcal{Y}$.

Without order information, every permutation of active proteins may lead to a different feasible network, yielding a combinatorial explosion of the feasible set. However, physical principles underlying signaling networks suggest that not all feasible solutions are equally likely, since closely connected proteins will co-occur more often. We exploit this intuition to devise a network inference strategy, explicitly modelling the sequential nature of signaling.

GNT is quite similar to the problem of *network inference from co-occurrences* (NICO) [1], which we previously developed to identify telecommunication networks [2]. The current paper investigates the GNT problem and extends NICO to noisy data, inevitable in biological experimentation, leading to new algorithms that are better suited to GNT. Thus, we refer to the mathematical problem tackled in this paper as *network inference from co-occurrences with uncertainty* (NICOU). The theory and methods herein presented are preliminary: much work needs to be done before GNT (potentially) becomes a useful tool for *in silico* genomics. The aim of this paper is to mathematically define the basic form of the GNT problem and to offer an initial solution scheme. For example, although it's suspected that some pathways are tree-structured or may include feedback loops, we limit our investigation to linear signaling pathways (simple protein-protein chains). Despite its limited scope, we hope this paper will stimulate further research in this important problem.

1.1. High-Throughput Sources of Co-Occurrence Data

There are several high-throughput genomic and proteomic measurement modalities that provide (possibly noisy) co-occurrence data.

1. Multiparameter flow cytometry (FC) simultaneously measures the activity levels of multiple proteins in a given sample. The FC data used in [3] only measures activity levels of 11 proteins, but novel labelling approaches promise to improve upon these limits, making FC applicable to larger protein network studies [4].

2. Gene knock-out experiments can also be used to construct a list of proteins involved in a pathway by including a protein in the list if its corresponding deletion mutant produces significantly different expression responses than the wildtype [5].

3. Protein chip data implicates proteins that are actively phosphorylated under a specific condition. Since phosphorylation reactions are a primary mechanism underlying cell signaling, this data can also be used to obtain lists of proteins in a pathway [6].

4. The existing literature provides a wealth of knowledge of putative signaling pathways and pathway components. In collaboration with researchers at the Stowers Institute and the University of Michigan, we have conducted GNT experiments using protein lists reported in the literature [7].

5. Protein-protein interactions (PPI) are a fundamental mechanism underlying cell signaling. Although PPIs do not directly indicate which pathways a protein belongs to, PPI data can easily be used as a probabilistic prior on network topology structure.

Note that these data types do not directly provide hard (binary) lists of proteins in each pathway. Rather, they may relate soft information such as the likelihood that each protein is active or a measure of the abundance of the protein. Our approach can accommodate this type of soft data, as well as (possibly noisy) binary data.

1.2. Related Work

Most previous work on inferring signaling pathways/networks from high-throughput data is based on learning the structure of a Bayesian network (probabilistic graphical model) [3, 8]. These approaches are based on the assumption that proteins with similar expression profiles are probably connected. In contrast, our approach is based on the intuition that proteins which co-occur frequently (*i.e.*, appear together in many pathways) are more likely to physically interact. The connectivity structure of graphical models corresponds to conditional dependency/independency relationships between protein activity levels, and it is recognized that this structure does not necessarily correspond to physical mechanisms underlying cell signaling [9]. Researchers have also proposed using additional experiments, after learning an initial graphical model, to recover *causal* network structure. For example, Sachs *et al.* [3] use FC data to learn a family of influence networks (directed graphical models–DGM); they then perform model averaging over this family of DGMs and use deletion interventions to resolve the resulting average DGM into a causal network. The solutions produced by their approach rely on the initial family of DGMs to capture all potential physical interactions.

Another body of work seeks to learn individual signaling pathways from gene expression and PPI data [10, 11, 12]. Such methods assume that the underlying network structure is given via PPI data and seek to learn the subset of edges in the network corresponding to a particular pathway. However, PPI is among the noisiest, least reliable high-throughput data; thus, the PPI networks, on which these procedures are based, may lack critical links or include extraneous ones. Rather than constraining the inferred network to be a subset of the noisy PPI network, we can incorporate it as a prior, allowing our procedure to tease out true connectivity using all of the available data.

2. AN INFERENCE STRATEGY FOR GNT

We adopt the following generative observation model. Let $\mathbf{w} = (w_1, \ldots, w_N) \in V^N$, be an ordered pathway. We model pathways as iid samples of a *Markov chain* (MC), parameterized by an initial state distribution, $p_i = \mathbb{P}(w_1 = i)$, for $i \in V$, and a matrix of transition probabilities, $P_{i,j} = \mathbb{P}(w_t = j | w_{t-1} = i)$, for $(i,j) \in V^2$, where $P_{i,j} > 0$ if and only if $(i,j) \in E$. This is not to say that biological systems actually work according to an MC; it is a simple mathematical model which aptly captures first-order network behavior. Let $\boldsymbol{\theta} = \{p_i\}_{i \in V} \cup \{P_{i,j}\}_{(i,j) \in V^2}$ denote the entire collection of MC parameters, which satisfy $\sum_{i=1}^{|V|} p_i = 1$, and $\sum_{j=1}^{|V|} P_{i,j} = 1$ for every $i \in V$. To model the fact that observations lack order information, we define a co-occurrence as a shuffled version of a path. Let $\pi : (1, \ldots, N) \to (\pi(1), \ldots, \pi(N))$ be a permutation drawn uniformly at random from $\mathbb{S}_N$, the set of all permutations of N objects. A co-occurrence $\mathbf{x} \in V^N$ is related to the path $\mathbf{w}$ via $w_t = x_{\pi(t)}$.

The discussion below will make use of equivalent binary representations for Markov chain samples, co-occurrences, and permutations. For a length-N path $\mathbf{w}$, the equivalent binary representation is a binary-valued matrix $\mathbf{W} \in \{0,1\}^{N \times |V|}$ such that if $w_t = i$ then $W_{t,i} = 1$ and $W_{t,j} = 0$ for all $j \neq i$; for a co-occurrence $\mathbf{x} \in V^N$, we have the analagous binary representation, $\mathbf{X} \in \{0,1\}^{N \times |V|}$. With this notation, the log-probability of a path can be written as

$$\log \mathbb{P}(\mathbf{W}|\boldsymbol{\theta}) = \sum_{i=1}^{|V|} W_{1,i} \log p_i + \sum_{t=2}^{N} \sum_{i,j=1}^{|V|} W_{t-1,i} W_{t,j} \log P_{i,j},$$

(with the usual convention $0 \log 0 = 0$). Let $\mathbf{A}^\pi \in \{0,1\}^{N \times N}$ be the permutation matrix corresponding to permutation π, that is, $A^\pi_{t,t'} = 1 \Leftrightarrow \pi(t) = t'$. In this notation, $\mathbf{W} = \mathbf{A}^\pi \mathbf{X}$, thus $W_{t,i} = \sum_{t'=1}^{N} A^\pi_{t,t'} X_{t',i}$, and we can write

$$\log \mathbb{P}(\mathbf{X}|\pi, \boldsymbol{\theta}) = \sum_{i=1}^{|V|} \sum_{t'=1}^{N} A^\pi_{1,t'} X_{t',i} \log p_i + \sum_{t=2}^{N} \sum_{i,j=1}^{|V|} \sum_{t',t''=1}^{N} A^\pi_{t-1,t'} X_{t',i} A^\pi_{t,t''} X_{t'',j} \log P_{i,j}. \quad (1)$$

Finally, we model a noisy observation of the co-occurrence $\mathbf{X}$ as a binary vector $\mathbf{y} \in \{0,1\}^{|V|}$, where $y_i = 1$ if vertex i is implicated in the path. We assume that the observations at different nodes (components of $\mathbf{y}$) are mutually independent, and that observation y_i only depends on whether node i actually occurs in the path. Let $S_i = \sum_{t=1}^{N} X_{t,i}$ denote a variable indicating whether vertex i occurs in this path; since each vertex appears at most once in a path, $S_i \in \{0,1\}$. We model y_i as an observation of S_i through a binary channel with (known) parameters $\rho_0 = \mathbb{P}(y_i = 0 | S_i = 0)$ and $\rho_1 = \mathbb{P}(y_i = 1 | S_i = 1)$. Denoting $\tilde{\rho}_1 = 1 - \rho_1$ and $\tilde{\rho}_0 = 1 - \rho_0$,

$$\mathbb{P}(y_i|S_i) = (\rho_1)^{y_i\, S_i}\ (\tilde{\rho}_1)^{(1-y_i)\, S_i}\ (\tilde{\rho}_0)^{y_i\, (1-S_i)}\ (\rho_0)^{(1-y_i)\, (1-S_i)}.$$

Finally, assuming mutual independence, $\mathbb{P}(\mathbf{y}|\mathbf{X}) = \prod_{i=1}^{|V|} \mathbb{P}(y_i|S_i)$.

2.1. Network Inference

Our goal is to infer the network structure E from a collection of M iid observations, $\mathcal{Y} = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}\}$. Given the model described above, network inference reduces to estimating the MC parameters $\boldsymbol{\theta}$. In particular, given an estimates of these parameters, we can determine the most likely structure of each path, and it is easy to determine the structure of a network from a set of ordered paths simply by inserting edges between adjacent vertices in each path.

Adopting the *maximum likelihood* (ML) criterion, we seek

$$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}}\ \log \mathbb{P}(\mathcal{Y}|\boldsymbol{\theta}), \quad (2)$$

subject to above mentioned constraints on $\boldsymbol{\theta}$. It is possible to write an exact expression for the objective function, $\log \mathbb{P}(\mathcal{Y}|\boldsymbol{\theta})$, based on the model described in the previous section. However, it turns out that, in general, the objective is a complicated, multi-modal, non-convex function of $\boldsymbol{\theta}$, so tools from convex optimization cannot be applied directly with any guarantees. Moreover, even computing $\log \mathbb{P}(\mathcal{Y}|\boldsymbol{\theta})$ at one particular value of $\boldsymbol{\theta}$ requires evaluating a number of terms which scales exponentially in both M and $|V|$. Thus, direct optimization methods (*e.g.*, gradient descent) are not computationally tractable in this setting. Determining the global optimum would require exhaustive search which is also computationally intractable.

3. EM ALGORITHM

We now describe an efficient *expectation-maximization* (EM) algorithm for solving (2), based on treating the co-occurrences, $\mathcal{X} =$

$\{\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}\}$, and permutations, $\Pi = \{\pi^{(1)}, \ldots, \pi^{(M)}\}$, as *missing data* which, together with the observations, $\mathcal{Y}$, form the *complete data*. The EM algorithm alternates between an E-step which computes

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^k) = \mathbb{E}\left[\log \mathbb{P}(\mathcal{Y}, \mathcal{X}, \Pi|\boldsymbol{\theta}) \;\middle|\; \mathcal{Y}, \boldsymbol{\theta}^k\right], \tag{3}$$

and an M-step, which updates the parameter estimate according to $\boldsymbol{\theta}^{k+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^k)$, subject to the constraints ensuring the parameters form a distribution. Because $\log \mathbb{P}(\mathcal{Y}|\boldsymbol{\theta})$ is continuous w.r.t. $\boldsymbol{\theta}$, and bounded above, standard convergence results for EM guarantee monotonic convergence to a local maximum [13].

3.1. Exact E-Step

Observe that, based on the assumption that observations are iid,

$$\log \mathbb{P}(\mathcal{Y}, \mathcal{X}, \Pi|\boldsymbol{\theta}) = \sum_{m=1}^{M} \log \mathbb{P}(\mathbf{y}^{(m)}, \mathbf{X}^{(m)}, \pi^{(m)}|\boldsymbol{\theta}). \tag{4}$$

Then, according to the model described in Section 2,

$$\begin{aligned}&\log \mathbb{P}(\mathbf{y}^{(m)}, \mathbf{X}^{(m)}, \pi^{(m)}|\boldsymbol{\theta})\\ &\quad = \log \mathbb{P}(\mathbf{y}^{(m)}|\mathbf{X}^{(m)}) + \log \mathbb{P}(\mathbf{X}^{(m)}|\pi^{(m)}, \boldsymbol{\theta}) + \log \mathbb{P}(\pi^{(m)}).\end{aligned}$$

Since $\log \mathbb{P}(\mathbf{y}^{(m)}|\mathbf{X}^{(m)})$ and $\log \mathbb{P}(\pi^{(m)})$ do not involve $\boldsymbol{\theta}$, they will not affect the outcome of the M-step, and so we can drop them.

Expression (1) shows that $\log \mathbb{P}(\mathbf{X}^{(m)}|\pi^{(m)}, \boldsymbol{\theta})$ is linear w.r.t. simple binary functions of $(\mathcal{X}, \Pi)$: $\alpha_i^{(m)} = \sum_{t'=1}^{N} A_{1,t'}^{\pi^{(m)}} X_{t',i}^{(m)}$, which equals one if and only if the mth path starts at vertex i, and

$$\beta_{i,j}^{(m)} = \sum_{t=2}^{N} \sum_{t',t''=1}^{N} A_{t-1,t'}^{\pi^{(m)}} X_{t',i}^{(m)} A_{t,t''}^{\pi^{(m)}} X_{t'',j}^{(m)},$$

which equals one if and only if the mth path contains a transition from i to j. Since the conditional expectation is a linear operator, given $\mathcal{Y}$ and $\boldsymbol{\theta}^k$, the E-step boils down to computing

$$\begin{aligned}\bar{\alpha}_i^{(m)} &= \mathbb{E}\left[\alpha_i^{(m)} \big| \mathbf{y}^{(m)}, \theta^k\right] \quad \text{for } i \in V \text{ and } m = 1, \ldots, M;\\ \bar{\beta}_{i,j}^{(m)} &= \mathbb{E}\left[\beta_{i,j}^{(m)} \big| \mathbf{y}^{(m)}, \theta^k\right] \quad \text{for } i, j \in V \text{ and } m = 1, \ldots, M.\end{aligned}$$

With these expected sufficient statistics in hand, we have

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}^k) \propto \sum_{m=1}^{M} \left(\sum_{i=1}^{|V|} \bar{\alpha}_i^{(m)} \log p_i + \sum_{i,j=1}^{|V|} \bar{\beta}_{i,j}^{(m)} \log P_{i,j} \right). \tag{5}$$

One can derive exact expressions for $\{\bar{\alpha}_i^{(m)}\}$, and $\{\bar{\beta}_{i,j}^{(m)}\}$. In fact, since the E-step decouples w.r.t. each observation, the resulting computational cost is linear in M (compare to direct optimization which is exponential in M). However, the need to marginalize over all possible co-occurrences makes the exact E-step have exponential cost in the number of vertices, $|V|$, thus being intractable except for very small networks. We will work around this issue in Section 4

3.2. M-Step

The M-step updates the parameter estimates, setting $\boldsymbol{\theta}^{k+1}$ equal to the maximizer of $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^k)$, which is a concave function of p_i and $P_{i,j}$. Maximizing (5), under the normalization constraints, yields

$$p_i^{k+1} = \frac{\sum_{m=1}^{M} \bar{\alpha}_i^{(m)}}{\sum_{i=1}^{|V|} \sum_{m=1}^{M} \bar{\alpha}_i^{(m)}} \tag{6}$$

$$P_{i,j}^{k+1} = \frac{\sum_{m=1}^{M} \bar{\beta}_{i,j}^{(m)}}{\sum_{j=1}^{|V|} \sum_{m=1}^{M} \bar{\beta}_{i,j}^{(m)}}. \tag{7}$$

4. MONTE CARLO EM VIA IMPORTANCE SAMPLING

Given the combinatorial nature of the exact E-step (enumerating every permutation of all co-occurrences), exact computation is generally intractable. Instead, we use Monte Carlo methods, *importance sampling* (IS), in particular, to approximate the sufficient statistics. To lighten notation in this section we drop the superscript k from $\boldsymbol{\theta}^k$, using $\boldsymbol{\theta}$, p_i, and $P_{i,j}$ to denote the current parameter estimates. Also, we focus on a particular path, thus drop the superscript (m).

Recall that $\bar{\alpha}_i$ and $\bar{\beta}_{i,j}$ are conditional expectations; ***e.g.***,

$$\begin{aligned}\bar{\alpha}_i &= \mathbb{E}\left[\sum_{t'=1}^{N} A_{1,t'}^{\pi} X_{t',i} \big| \mathbf{y}, \boldsymbol{\theta}\right]\\ &= \sum_{\mathbf{X}} \sum_{\pi} \left(\sum_{t'=1}^{N} A_{1,t'}^{\pi} X_{t',i} \right) \mathbb{P}(\mathbf{X}, \pi|\mathbf{y}, \boldsymbol{\theta}).\end{aligned} \tag{8}$$

An ideal Monte Carlo algorithm would generate L iid sample pairs $\{(\mathbf{X}^\ell, \pi^\ell)\}_{\ell=1}^{L}$ from the joint distribution $\mathbb{P}(\mathbf{X}, \pi|\mathbf{y}, \boldsymbol{\theta})$, and use them to approximate this expectation. However, this is not practical in the present setting since determining the distribution $\mathbb{P}(\mathbf{X}, \pi|\mathbf{y}, \boldsymbol{\theta})$ requires evaluating it for every possible co-occurrence and permutation, which is the task we are trying to avoid in the first place. Instead we propose the following sequential scheme for sampling co-occurrences and permutations.

First, we sample a co-occurrence by drawing S_i^ℓ from $\{0, 1\}$ independently for each $i \in V$, such that $S_i^\ell = 1$ with probability $y_i\, \rho_1 + (1 - y_i)\tilde{\rho}_0$; this produces a co-occurrence $\mathbf{X}^\ell$.

Next, we sample a permutation, π^ℓ, given $\mathbf{X}^\ell$, using the sequential procedure developed in [1]. This amounts to sampling a probable ordering of the vertices in $\mathbf{X}^\ell$. To ensure that no co-occurring vertex is sampled twice we use a vector of binary flags, $\mathbf{f} \in \{0, 1\}^{|V|}$. Given a probability distribution $\mathbf{p} = (p_1, \ldots, p_{|V|})$ on the vertex set, V, denote by $\mathbf{p}|\mathbf{f}$ the restriction of $\mathbf{p}$ to $\mathbf{f}$, ***i.e.***,

$$(\mathbf{p}|\mathbf{f})_i = \frac{p_i f_i}{\sum_{j=1}^{|V|} p_j f_j}; \quad \text{for } i \in V. \tag{9}$$

We sample a permutation as follows:

Step 1: Initialize $\mathbf{f}$ so that $f_i = S_i^\ell$ for all $i \in V$. Sample an element v from V according to the distribution $\mathbf{p}|\mathbf{f}$ on V, where $\mathbf{p} = (p_1, \ldots, p_{|V|})$ is the initial state distribution. Find t such that $X_{t,v}^\ell = 1$ and set $\pi(1) \leftarrow t$. Update $f_v \leftarrow 0$ and set $i \leftarrow 2$.

Step 2: Let $\mathbf{P}_v = (P_{v,1}, \ldots, P_{v,|V|})$, ***i.e.,*** the vth row of transition probabilities. Sample an element v' according to the distribution $\mathbf{P}_v|\mathbf{f}$. Find t' such that $X_{t',v'}^\ell = 1$. Set $\pi(i) \leftarrow t'$ and $f_{v'} \leftarrow 0$.

Step 3: If $i < \sum_{i=1}^{|V|} S_i^\ell$, update $v \leftarrow v'$ and $i \leftarrow i + 1$, and go to Step 2; otherwise stop.

Repeating this sampling procedure L times yields a collection of iid samples $\{(\mathbf{X}^\ell, \pi^\ell)\}_{\ell=1}^{L}$, where the superscript now identifies

the sample number. Samples generated according to this scheme are drawn from a distribution $\mathbb{Q}(\mathbf{X}, \pi|\mathbf{y}, \boldsymbol{\theta})$ which is different from the distribution $\mathbb{P}(\mathbf{X}, \pi|\mathbf{y}, \boldsymbol{\theta})$. IS estimates correct for this disparity and are given by the expressions

$$\begin{aligned}\widehat{\alpha}_i &= \frac{\sum_{\ell=1}^{L} u_\ell \sum_{t'=1}^{N} A_{1,t'}^{\pi^\ell} X_{t',i}^{\ell}}{\sum_{\ell=1}^{L} u_\ell} \\ \widehat{\beta}_{i,j} &= \frac{\sum_{\ell=1}^{L} u_\ell \sum_{t=2}^{N} \sum_{t',t''=1}^{N} A_{t-1,t'}^{\pi^\ell} X_{t',i}^{\ell} A_{t,t''}^{\pi^\ell} X_{t'',j}^{\ell}}{\sum_{\ell=1}^{L} u_\ell},\end{aligned}$$

where the correction factor u_ℓ is given by

$$u_l = \frac{\mathbb{P}(\mathbf{X}^\ell, \pi^\ell|\mathbf{y}, \boldsymbol{\theta})}{\mathbb{Q}(\mathbf{X}^\ell, \pi^\ell|\mathbf{y}, \boldsymbol{\theta})} = \prod_{t=2}^{N} \sum_{t'=t}^{N} P_{x_{\pi^\ell(t-1)}, x_{\pi^\ell(t)}}. \qquad (10)$$

Correcting in this fashion ensures that the importance sample estimates are consistent. In fact, terms in the product (10) are byproducts of Step 2 of the sampling algorithm (denominators of $(\mathbf{P}_v|\mathbf{f})$).

When using IS in place of the exact E-step, one is confronted with the question: *what number of samples, L, is sufficient?* Without enough samples the IS approximations will be of poor quality, and the EM will no longer converge to a local maximum because it is overwhelmed with sampling error. On the other hand, our motivation for using IS in the first place is because the exact E-step is intractable, so we would like to use as few samples as possible while still guaranteeing convergence. It is possible to show (using arguments similar to those in [1]) that the IS-based EM keeps the monotonic convergence properties of the exact EM when using a number of samples, L, which grows polynomially in both the network size $|V|$, and the number of observations, M. In particular, each IS-based EM iteration monotonically increases the objective function with probability at least $1 - \delta(L)$, where $\delta(L)$ tends to 0 exponentially as $L \to \infty$. We omit the details due to lack of space.

5. SIMULATIONS

This section reports the outcome of testing our algorithm on simulated data. The observations are noisy co-occurrences sampled from a random network with 50 vertices (intended to reflect the size and complexity of currently known/conjectured signal transduction networks). We generate $M = 50$, 100, and 200 paths, by randomly choosing sources and destination vertices, and randomly generating paths between them. The paths are reduced to co-occurrences, and then observations are obtained by contaminating the co-occurrences with binary noise (as described above) with $\rho_0 = \rho_1 = 0.01$. We compare the performance of the NICOU scheme introduced in this paper, with that of the NICO algorithm described in [1] which does not account for observation uncertainty. Performance is measured in terms of the normalized ℓ_1 error $\epsilon_1(\widehat{\boldsymbol{\theta}}) = \frac{1}{|V|} \sum_{i,j=1}^{|V|} \left|P_{i,j}^* - \widehat{P}_{i,j}\right|$, where $\{P_{i,j}^*\}$ are the transition probabilities corresponding to the true paths underlying the observations, and $\{\widehat{P}_{i,j}\}$ are the estimated transition probabilities. This metric measures how well the inferred parameters match what one would be able to recover if it were possible to precisely measure ordered paths through the network.

Table 1 summarizes the results of our experiments. When only $M = 50$ observations are made, there is not enough "signal", relative to the amount of noise, to tease apart true network structure from errors in the observations, and both techniques do equally as well. However, as more and more measurements become available, NICOU performs significantly better than NICO, as it is able to compensate for extraneous or missing elements in each pathway.

M	NICO	NICOU
50	0.5331	0.5688
100	0.4030	0.3196
200	0.0875	0.0405

Table 1. Simulation results. Comparing the average ℓ_1 error of the NICOU algorithm described in this paper with the NICO algorithm of [1], which does not account for uncertainty in the measurements. Each entry represents the average over 10 random networks.

6. CONCLUSION

This paper poses the problem of genomic network tomography and proposes an initial solution. Our ongoing work involves extending the algorithm to handle more complicated pathway measurements (trees and other graphs) and other types of biological data. We are also beginning the process of testing NICOU on real biological data. Based on the initial success reported in [7], we believe that the proposed approach to genomic network tomography is very promising.

7. REFERENCES

[1] M. Rabbat, M. Figueiredo, and R. Nowak, "Network inference from co-occurrences," submitted to *IEEE Transactions on Information Theory*, May 2006.

[2] M. Rabbat, J. Treichler, S. Wood, and M. Larimore, "Understanding the topology of a telephone network via internally-sensed network tomography," in *Proc. IEEE ICASSP*, vol. 3, pp. 977–980, Philadelphia, PA, 2005,

[3] K. Sachs, O. Perez, D. Pe'er, D. Lauffenburger, and G. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," *Science*, vol. 308, pp. 523–29, 2005.

[4] P. Krutzik and G. Nolan, "Fluorescent cell barcoding in flow cytometry allows high-throughput drug screening and signaling profiling," *Nature Methods*, vol. 3, pp. 361–68, 2006.

[5] G. Giaever *et al.*, "Functional profiling of the *s. cerevisae* genome," *Nature*, vol. 418, no. 6896, pp. 387–91, July 2002.

[6] J. Ptacek *et al.*, "Global analysis of protein phosphorylation in yeast," *Nature*, vol. 438, no. 7068, pp. 679–84, Dec. 2005.

[7] D. Zhu, M. Rabbat, A. Hero, R. Nowak, and M. Figueiredo, "De novo signaling pathway reconstruction from multiple data sources," in *New Research in Signal Transduction*. 2006, Nova.

[8] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using bayesian networks to analyze expression data," *J. Computational Biology*, vol. 7, pp. 601–620, 2000.

[9] T. Gardner and J. Faith, "Reverse-engineering transcription control networks," *Phys. of Life Rev.*, vol. 2, pp. 65–88, 2005.

[10] M. Steffen, A. Petti, J. Aach, P. D'haeseleer, and G. Church, "Automated modelling of signal transduction networks," *BMC Bioinformatics*, vol. 3, pp. 34, November 2002.

[11] C.-H. Yeang, T. Ideker, and T. Jaakkola, "Physical network models," *J. Comp. Biology*, vol. 11, no. 2-3, pp. 243–62, 2004.

[12] Z. Tu, L. Wang, M. Arbeitman, T. Chen, and F. Sun, "An integrative approach for causal gene identification and gene regulatory pathway inference," *Bioinformatics*, vol. 22, no. 14, pp. e498–96, 2006.

[13] C. Wu, "On the convergence properties of the EM algorithm," *Ann. of Statistics*, vol. 11, no. 1, pp. 95–103, 1983.